利用深度学习研究中文书写系统、字体对阅读绩效的影响

张雨馨 刘斌* 黄思奇
西南财经大学 统计学院 统计研究中心 成都 610031

摘要：
[目的] 本文针对中文出版物中不同字体、不同书写系统的阅读绩效进行客观对比研究。
[方法] 具体地，将汉字渲染成其对应字形的图像，并进一步按照语序把句子中的汉字图像折叠成为三维的句子张量。对于同一段中文文本，用不同的字体或者简体、繁体会得到句子的视觉差异化的张量表达。通过进一步将得到句子张量输入到我们设计的深度语言模型，进行文本分类等任务的测试，可以客观地比较字体和书写系统对阅读绩效的影响。
[结果] 通过在两个中文文本分类数据集上的测试发现，一些特殊不常用字体相较于常用字体的机器识别准确度较低，并且常用字体中不同字体的阅读绩效也有差异。
[结论] 通过假设检验得出使用楷体和黑体的数据集在文本分类任务上的准确度存在显著性差异，楷体相比于黑体来说阅读绩效更高。简体中文和繁体中文的阅读绩效存在显著性差异。
关键词： 阅读绩效；文本分类；深度神经网络
分类号： TP391

Study How Chinese Writing System Chinese Fonts Affect Reading Performance Using Deep Learning

Zhang Yuxin[1], Huang Siqi[1] Liu Bin*
Center of Statistics Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu 610031

Abstract:
[Objective] We study the reading performance of different fonts and writing systems that are using in Chinese publications.
[Methods] Specifically, the Chinese characters in a sentence are rendered into their corresponding glyph images, then fold those images into a three-dimensional sentence tensor according to the word order. For different fonts or simplified/traditional Chinese text, we can get the corresponding representations with visual differences. By inputting the obtained sentence tensor into the proposed deep language model, we test them on text classification, which can objectively study the influence of font and writing system on reading performance.
[Results] According to the experiments on two real-world Chinese text classification datasets, Toutiao and Thucnews, we found that the accuracy of text classification on some uncommon fonts is lower than that of common used fonts, and the text representation efficiency of different fonts in the common fonts is also different.
[Conclusions] Through a hypothesis test, we found that there is a significant difference in

通信作者： 刘斌（liubin@swufe.edu.cn）*
论文第一、二作者为在读硕士研究生

the accuracy of using the data sets of regular script and bold script for text classification task, and the efficiency of regular script is higher than that of bold script. There are significant differences in reading performance between simplified and traditional writing systems.
Keywords: reading performance; text classification; deep neural network

中文文字的不同字形特征在阅读中会产生不同的视觉心理效果[1]，因而会带来阅读绩效上的差异。比如，大部分的中文印刷品、网页的正文多选用楷体或宋体，而较少用黑体[2]。对于中文而言，除了字体带来的视觉心理差异[3]，并存的两种书写系统，简体中文和繁体中文，同样会给文字表达效率带来视觉心理上的差异[4]。对于整个汉字体系发展来说，有些简体字看似没有繁体字的表意属性强，但这并不表明简体字推行后汉字的表意属性就降低了[5]。如，“塵”简化成“尘”、“滅”简化成“灭”后，简化字比繁体字更好理解。关于简体和繁体的优劣之争的研究由来已久[4,5,6]。与此同时，学界部分学者力求探讨汉字字体的视觉传达与识别性能[7]。从上世纪 20、30 年代起，汉字实验研究一直都是语言学界、心理学界共同关注的课题。语言学家对汉字的各种识别研究多从文字学角度入手[8]，如汉字结构特征与汉字识别相关性等课题。另有一批专家立足于汉字信息学及认知心理学相关知识[9]，对人类识别汉字信息符号的心理过程进行仿真模拟，用以解决汉字计算机信息处理识别系统中存在的各种现实问题。除此之外，一些学者从生理学[10,11]、心理学[12]角度出发，对汉字不同字体的辨别率做实验性研究。如，金文雄等[12]以判读正确率为指标，在三种照明条件下，比较了宋体、黑体、长仿宋体和正仿宋体四种汉字字体的阅读绩效。

已有的关于汉字字体识别的研究，多来自于与文字相关的社会学领域，基本是从语言文字学、心理学等角度出发，其研究的方法因学科性质大多较为主观。本文试图用机器阅读来客观地研究中文字体、书写系统和文本阅读绩效的关系。自然阅读的阅读绩效可以用机器文本分类的准确率来近似，后者是前者在机器阅读上的同等概念。

具体地，我们利用文字字形来作为文本的表达，然后输入自然语言模型来进行文本分类测试。即，将句子的单词或字符渲染成图像，然后将它们折叠成三维句子张量 $\chi \in \mathbb{R}^{w\times h\times l}$，其中 w 是单词或字符图像的大小，l 是句子的长度, 如图 1 所示。每个切片 $\chi_i \in \mathbb{R}^{w\times h}$对应一个单词或一个字符图像。因此，句子可以被一个三维句子张量 χ 来表达。进而，我们把同一段文本按照不同字体或者不同书写系统表达成对应的句子张量的形式，并在机器语言模型下测试其识别性能。本文中，基于句子的张量表达，利用 Liu 等提出的三维卷积语言模型[13]来近似计算阅读绩效。

本文的主要贡献包括 2 个方面:

1) 从机器阅读的角度，提出一种客观评价中文简体、繁体等不同书写系统，以及不同字体的阅读绩效的方式;

2) 利用假设检验，在头条数据集和清华数据集上验证了不同书写系统、不同字体对阅读绩效的影响。

通过在两个中文文本分类数据集上的测试发现，常用字体和不常用字体的阅读绩效存在差异。一些常用字体，比如楷体和黑体的阅读绩效也存在显著差异。简体和繁体两个书写系统的阅读绩效具有显著性差异。

1 研究方法

1.1 概论

首先我们将一个句子 S 渲染成一个如图 1 所示的三维的张量，其中这个张量的每一个切

片对应于一个汉字，即 S 中的汉字v_i被渲染为图像$\chi_i \in \mathbb{R}^{w\times h}$。然后将 S 中每个字按顺序折叠成三维句子张量 $\chi \in \mathbb{R}^{w\times h\times l}$，其中 w 是单词或字符图像的大小，l是句子的长度。我们将大小为 $w\times h\times n$ 的三维卷积核应用于“文本张量”，其中 w 和 h 分别是字符图像的宽度和高度，n 是字符数。换言之，3D 卷积一次滑动作用于 n 个字符，相当于一次 n-gram 的特征检测。我们可以通过改变 n 的值，得到不同大小的 n-gram 检测器，并可以使用多个 n-gram 卷积来提取文本特征。例如，在我们的实验中，n 可以取{2，3，4，5}的值。在我们所提出的框架下，多个 n-gram 的集成可以非常容易和快速地实现。

在神经网络语言建模中，正序和反序的文本信息是两种不同的输入[13]。我们采用双向卷积来提取文本特征。



Fig. 1 Three-dimensional sentence tensor

图 1 三维句子张量

1.2 网络搭建

神经网络的体系结构可以描述如下[13]：

1. **三维卷积层**: *kernel size = (20,20，2), stride = (1, 1, 1), number of kernels = 50, padding = 0*;
   kernel size = (20,20,3), stride = (1, 1, 1), number of kernels = 50, padding = 0;
   kernel size = (20,20,4), stride = (1, 1, 1), number of kernels = 50, padding = 0;
   kernel size = (20,20,5), stride = (1, 1, 1), number of kernels = 50, padding = 0;
2. **MaxPool1d 层** (the max-over-time pooling): *kernel size = 3, stride = 3, dilation = 2, padding = 0*;
3. **全连接层 1**: *input = 100*99, output =1100*;
4. **全连接层 2**: *input =1100, output = 120*;
5. **全连接层 3**: *input = 120, output = 类别数量.*

2 实验过程

2.1 实验配置

根据现有的数据集，我们从文本分类的任务上来测试不同字体和简体中文、繁体中文的分类准确率。这里分类准确率可以近似等价于阅读绩效。

数据集如表 1 所示。THUCNews 数据集[14]是根据 2005 年至 2011 年从新浪新闻 RSS[1]订阅渠道获得的历史数据生成的。原始的清华数据集是 14 分类，但是个别分类样本太少，为了避免出现样本不均衡问题，我们删除掉样本比较少的 4 个类别，最终我们使用的数据集为 10 类。Toutiao 新闻数据集从 Toutiao App 收集文本。每个项目包含新闻的标题和关键词。数据处理参照文献[13] 。对于所用的两个数据集，表 1 中给出了训练、验证和测试的样本大小。

[1] https://rss.sina.com.cn/

Table 1 Splitting of the sample size for training,validation and testing of the datasets for text classification.

表 1 用于训练、验证和测试文本分类数据集的样本大小的拆分

Datasets	*Training*	*Validation*	*Testing*	*Classes*	*Average length*	*Content*
THUCNews	16000	2000	2000	10	251	News
Toutiao	266318	37666	76471	15	38	Title and keywords

2.2 不同字体之间假设检验

我们将流行出版物中使用频率高的字体列为常用字体。常用字体来源于 Windows 自带字体库，不常用字体来源于方正字库[2]。

常用字体：宋体(SIMSUN)，楷体(STKAITI)，黑体(SIMHEI)，等线(DENG)，华文仿宋(STFANGSO)。

不常用字体：方正舒体(FZSTK)，方正字迹-长江行书简体(FZZJ-CJXSJW)，方正大草简体(FZDCJW)，方正鲁迅简体(FZLUXTJW)，方正字迹-欧阳长迪行楷(FZZJ-OYCDXKJW)。

2.2.1 常用字体和不常用字体检验

我们将这两个数据集渲染成 5 种常用字体和 5 种不常用字体，然后在它们上运行 *m* 次文本分类任务。从表 2 中的结果可以看出，常用字体和不常用字体有不同的形式，它们在阅读绩效上也有一定的差异。

学而时习之，不亦乐乎？
学而时习之，不亦乐乎？
学而时习之，不亦乐乎？
学而时习之，不亦乐乎？
学而时习之，不亦乐乎？
学而时习之，不亦乐乎？

Fig. 2 SIMSUN、STKAITI、SIMHEI、FZZJ-CJXSJW、FZLUXTJW、FZZJ-OYCDXKJW

图 2 宋体、楷体、黑体、方正字迹-长江行书简体、方正鲁迅简体、方正字迹-欧阳长迪行楷的视觉效果展示

Table 2 The results of text classification task for data sets with different fonts

表 2 对不同字体的数据集进行文本分类任务结果展示

Fonts	*Toutiao (m=40)*	*THUCNews(m=40)*
STKAITI 楷体	0.85075	0.9325
STFANGSO 华文仿宋	0.85075	0.93125

[2] https://www.foundertype.com/index.php/FindFont/index

SIMHEI 黑体	0.84625	0.92975
SIMSUN 宋体	0.8495	0.93025
DENG 等线	0.8495	0.93000
FZDCJW 方正大草简体	0.84775	0.9315
FZSTK 方正舒体	0.8465	0.93000
FZLUXTJW 方正鲁迅简体	0.84975	0.93075
FZZJ-OYCDXKJW 方正字迹-欧阳长迪行楷	0.84375	0.92925
FZZJ-CJXSJW 方正字迹-长江行书简体	0.8395	0.92875

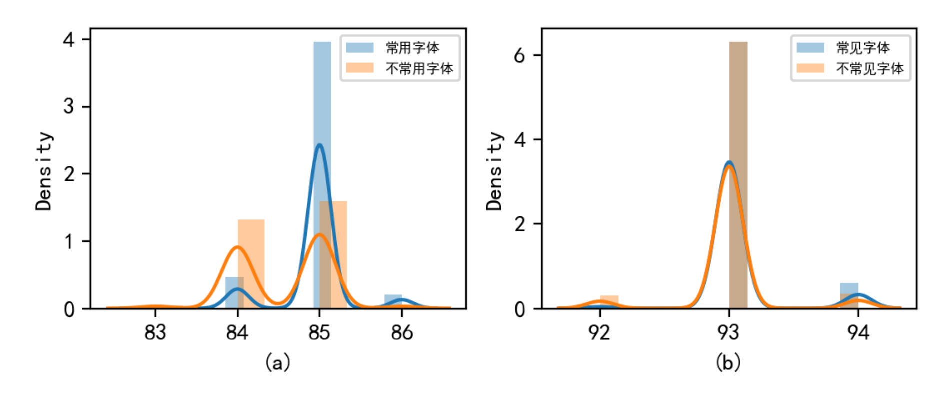


Fig. 3 The classification and recognition accuracy distribution of common fonts and uncommon fonts, where (a) is the Toutiao dataset and (b) is the Thucnews dataset.

图 3 常用字体和不常用字体的分类识别准确率分布，其中 (a)是头条数据集，(b)是清华数据集。

Table 3 The average classification and recognition accuracy of five commonly used fonts and five rarely used fonts

表 3 五种常用字体分类识别准确率均值和五种不常用字体分类识别准确率均值

Fonts	*Toutiao(m=200)*	*Thucnews(m=200)*
Common	0.84935	0.93075
Uncommon	0.84545	0.93005

图 3 展示了在头条数据集和清华数据集上使用常用字体和不常用字体进行分类识别任务的准确率分布。对常用字体和不常用字体阅读绩效的假设检验结果如表 4 所示：

Table 4 Hypothesis test on reading performance of data sets using common and uncommon fonts

表 4 对使用常用字体和不常用字体的数据集阅读绩效的假设检验

dataset	*T 值*	*自由度*	*P 值*
toutiao	8.036	362.275	1.3266e-14
thucnews	2.300	397.633	0.0220

H_0：机器对使用常用字体和不常用字体的阅读绩效无显著性差异

头条数据集检验统计量 t=8.036 ，自由度为 362.275，双尾检验 p 值= 1.3266e-14。在显著性水平设为 0.05 下，p 值小于显著性水平 **α**，所以拒绝原假设，有统计显著，即模型对使用常用字体和不常用字体的头条数据集的阅读绩效存在显著差异。

清华数据集检验统计量 t 值=2.300 自由度为 397.633 双尾检验 p 值= 0.0220。在显著性水平设为 0.05 下，p 值小于显著性水平 **α**，所以拒绝原假设，有统计显著，即模型对使用常用字体和不常用字体的清华数据集阅读绩效存在显著差异。

根据以上两个数据集得出的结论，常用字体与不常用字体之间的阅读绩效是存在显著差异的，也就是说不同字体的阅读绩效确实存在显著差异。这就很好的解释说明了为什么很多印刷书籍是有字体的常用选择范围的。阅读绩效高的的字体能帮助人们阅读，给人们更好的阅读体验，而不适宜的字体适得其反，严重妨碍阅读也会影响图书的销量。

2.2.2 常用字体间的检验

宋体、黑体、楷体、仿宋、等线虽然都作为基本常用字体，但依然存在阅读绩效的差异。接下来我们对常用字体的阅读绩效排序，在表 5 中可以看到，在头条数据集中，阅读绩效最高的字体为楷体，最低的字体为黑体；在清华数据集中，阅读绩效最高的字体为楷体和华文仿宋，最低的字体为黑体。我们进一步检验使用楷体和黑体的两个数据集阅读绩效有无显著差异。

学	而	时	习	之	，		不	亦	乐	乎	？
学	而	时	习	之	，		不	亦	乐	乎	？

Fig. 4 Hypothesis tests are presented in bold and regular fonts

图 4 假设检验所用黑体和楷体字体展示

Table 5 shows the results of text classification of datasets with different common fonts

表 5 对使用不同常用字体的数据集文本分类的结果展示

Common fonts	*Toutiao(m=40)*	*Thucnews(m=40)*
STKAITI	0.85075	0.9325
STFANGSO	0.85075	0.93125
simsun	0.8495	0.93025
DENG	0.8495	0.93
SIMHEI	0.84625	0.92975

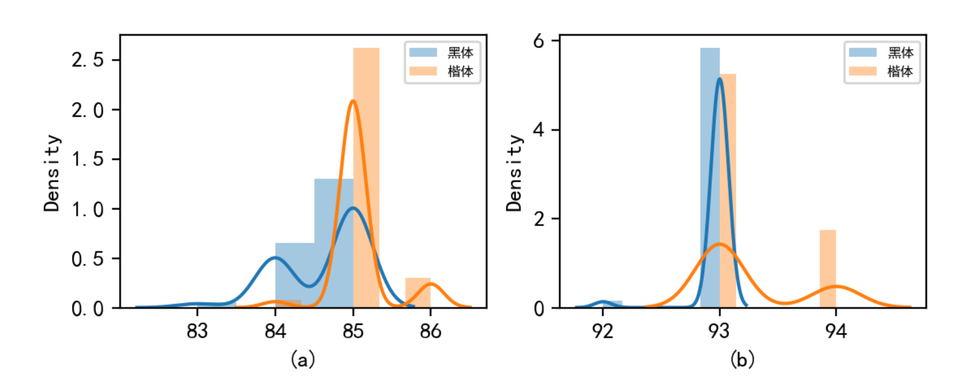


Fig. 5 The classification and recognition accuracy distribution of bold and regular type, where (a) is the Toutiao dataset and (b) is the Thucnews dataset.

图 5 黑体和楷体的分类识别准确率分布，其中 (a)是头条数据集，(b)是清华数据集。

Table 6 Hypothesis tests on reading performance of feature rendering with regular and bold characters

表 6 模型对使用楷体和黑体的数据集阅读绩效的假设检验

dataset	*T 值*	*自由度*	*P 值*
Toutiao	-4.4227	66.837	3.6756e-05
Thucnews	-3.7310	48.971	0.0005

图 5 展示了在头条数据集和清华数据集上使用黑体和楷体进行分类识别任务的准确率分布。对黑体和楷体阅读绩效的假设检验结果如表 6 所示.

H_0：机器对使用楷体和黑体的阅读绩效无显著性差异

头条数据集检验统计量 t 值= -4.4227，自由度为 66.83,双尾检验 p 值= 3.6756e-05。在显著性水平设为 0.05 下，p 值小于显著性水平 **α**，所以拒绝原假设，有统计显著，即模型对使用楷体和黑体的头条数据集的阅读绩效存在显著性差异。

清华数据集检验统计量 t 值= -3.7310，自由度为 48.971，双尾检验 p 值=0.0005。在显著性水平设为 0.05 下，p 值小于显著性水平 **α**，所以拒绝原假设，有统计显著，即模型对使用楷体和黑体的清华数据集的阅读绩效存在显著性差异。

我们以上实验可以得出结论，使用楷体和黑体的数据集的阅读绩效存在显著性差异，楷体的阅读绩效相对来说比黑体高些。我们猜想从视觉上来看，楷体结构部位之间互不连接，字体清楚，而黑体横竖的笔形粗细是相等的，文章如果通篇采用黑体给读者阅读，每一个字都非常醒目，会使读者产生视觉疲劳，而楷体相比于黑体，笔画没有那么粗壮，视觉上给人直观清楚的效果，所以这也符合我们黑体一般用于标题，楷体一般用于正文的排版习惯，较易于人们阅读。

2.3 简体中文和繁体中文假设检验

中文存在两种写作系统，即简体中文和繁体中文。几乎所有的汉语方言都是基于这两个写作系统，一般繁体中文比简体中文有更多的笔画。多年来，人们对繁体中文和简体中文进行了广泛的争论，例如，繁体中文和简体中文的区别，哪一个更有效率等等。

这一部分我们在模型框架下比较了简体中文和繁体中文的阅读绩效差异。我们在头条数据集和清华数据集上分别使用简体中文和繁体中文，然后运行 40 次文本分类任务。

学	而	时	习	之	，	不	亦	乐	乎	?
學	而	時	習	之	，	不	亦	樂	乎	?

Fig. 6 Simplified and Traditional Chinese Fonts

图 6 简体和繁体字体展示

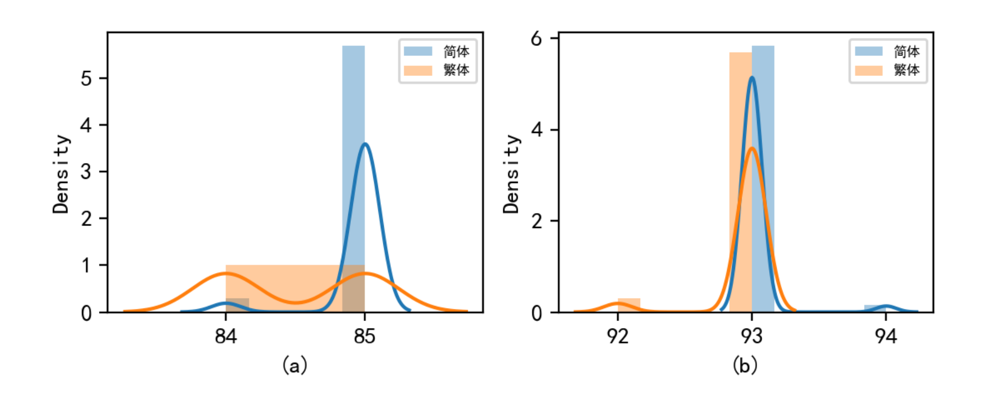


Fig. 7 The classification and recognition accuracy distribution of simplified and traditional characters, where (a) is the Toutiao dataset and (b) is the Thucnews dataset.

图 7 简体和繁体的分类识别准确率分布，其中 (a)是头条数据集，(b)是清华数据集。

Table 7 Text expression efficiency of Simplified Chinese and Traditional Chinese on different data sets

表 7 简体中文和繁体中文在不同数据集上的阅读绩效

Chinese writing system	*Toutiao(m=40)*	*Thucnews(m=40)*
Simplified Chinese	0.8495	0.93025
Traditional Chinese	0.8450	0.9295

图 7 展示了在头条数据集和清华数据集上使用简体和繁体进行分类识别任务的准确率分布。对简体和繁体阅读绩效的假设检验结果如表 8 所示：

Table 8 Hypothesis test on reading performance of data sets using simplified and traditional characters

表 8 模型对使用简体和繁体的数据集阅读绩效的假设检验

dataset	*T 值*	*自由度*	*P 值*
Toutiao	5.152	53.304	3.8302e-06
Thucnews	1.747	70.683	0.0850

H_0：机器对使用简体和繁体的阅读绩效无显著性差异

头条数据集检验统计量 t 值=5.152，自由度为 53.304，双尾检验 p 值= 3.8302e-06，在显著性水平设为 0.05 下，p 值小于显著性水平 **α**，所以拒绝原假设，有统计显著，即模型对使用简体和繁体的头条数据集的阅读绩效存在显著性差异。

清华数据集检验统计量 t 值=1.747，自由度为 70.683，双尾检验 p 值= 0.0850，在显著性水平设为 0.05 下，p 值大于显著性水平 **α**，所以不拒绝原假设，无统计显著，即没有充足的理由证明模型对使用简体和繁体的清华数据集的阅读绩效存在显著性差异。

简体和繁体两个书写系统，繁体字注重表意，每个字都有渊源，每个字意都有详细的演化过程，字形和字义相融匹配。而简体字笔画减少，一字多义情况也时常发生，这对模型的识别效率也有一定的影响。在头条数据集上的结果显示繁体和简体的阅读绩效存在显著差异，而在清华数据集上结果显示它们的阅读绩效不存在显著差异。我们猜想是因为头条数据

集样本比较短，准确率计算起来更加精细，所以头条数据集准确率都比清华数据集低，而清华数据集样本较长，所以对使用简体中文和繁体中文的数据集阅读绩效差异不大。由于在头条数据集上假设检验的结果已经有充足的理由拒绝无显著差异的原假设，所以我们认为，模型对使用繁体和简体的数据集阅读绩效存在显著性差异。

3 结论

通过句子张量，句子中文本特征都可以通过多个 *n*-gram 以正常顺序和反向顺序提取。为了研究中文不同字体和不同书写系统的识别差异，我们在几个数据集上进行文本分类任务。最后，我们根据在不同书写系统、不同字体下数据集文本分类的准确率做假设检验，假设检验结果为常用字体和不常用字体的阅读绩效存在差异。常用字体中楷体和黑体的阅读绩效也存在显著性差异。简体和繁体两个书写系统的阅读绩效存在显著性差异。本研究结果证实了书面出版物使用的字体存在阅读绩效的差异。但本研究使用的不常用字体仅来源于方正字库，其他不常用字体的表达效率还需进一步研究。

参 考 文 献

[1] 丁晓青.汉字识别研究的回顾[J].电子学报,2002,30(9):1364-1368

[2] 傅永和.浅析四种印刷字体[J].语文建设,1993(5):24-26

[3] 王丽佳.书籍文字设计的视觉心理效应[J].黑龙江生态工程职业学院学报,2007(01):124-125

[4] 黄敏聪.基于人文计算的汉字简繁体演变定量分析[J].科技视界,2012,18:64-67

[5] 谢金良.关于繁体字与简体字的若干思考[J].闽江学院学报,2009(30):45-49

[6] 胡志明.汉字发展史上的繁简字[J].海峡教育研究,2016(02):43-47

[7] 吴双翼.从汉字构造方法看视觉传达设计的意象关系[J].当代艺术,2011,000(002):95-96

[8] 万业馨.从汉字识别谈汉字与汉字认知的综合研究[J].语言教学与研究,2003(2):72-79

[9] 管益杰,方富熹.我国汉字识别研究的新进展[J].心理科学进展,2000,8(2):1-6

[10] 信伟, 陶道典, 黄涛,等.阅读中文简体字与繁体字人眼短暂性近视量及其回退时间的差异[J].中华眼视光学与视觉科学杂志, 2012, 14(10):581-583

[11] 信伟, 沈品呈, 郑晨琛,等.正视眼与近视眼阅读简体字与繁体字的调节微波动差异研究[J].中华眼科杂志, 2018(4):288-293

[12] 金文雄,朱祖祥,沈模卫.汉字字体对判读效果的影响[J].应用心理学,1992,7(3):8-11

[13] Bin Liu, Liang Wang, Guosheng Yin. Learning distributed sentence vectors with bi-directional 3D convolutions[C]// Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online):International Committee on Computational Linguistics,2020:6820-6830

[14] Jingyang Li, Maosong Sun. Scalable term selection for text categorization. [C]//Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL),Prague,Czech Republic:Association for Computational Linguistics,2007:774-782